CHAPTER 2: SUPPLEMENTAL FIELD TEST DATA ANALYSIS

Introduction

The June 30 Year 1 Evaluation Report included extensive analyses of the HSEE multiple choice (MC) questions for mathematics and English/language arts (ELA) in the Spring 2000 field test. At that time, hand-scoring of answers to the ELA essay questions[1] had not been completed. Scores for the essay questions were received on June 28. In this section, we describe the results of our analyses of these questions. First, we examined statistical indicators of the functioning and quality of the scores for the essay questions. Next, we analyzed the consistency of the scores of the responses provided by alternative raters. Then we turned to consideration of the effect of adding the essay questions on the accuracy of overall student scores. We conclude this section with additional analyses of the degree to which the field test samples are representative of California 10th grade students in general.

Quality of the Essay Questions

Booklet Design. After completing 100 multiple-choice questions, each student in the English Language Arts field test was presented two prompts, each requiring an essay response. Responses to the first prompt were scored twice, once as a reading measure based on the content of the response and a second time as a writing measure based on mechanics and style. Three versions of each of the four field test booklets were created, with a different pair of essay questions included in each version of each booklet. The Spring 2000 field test thus included 12 pairs of essay questions.

Scoring. Each response was read, independently, by two different scorers. Following the scoring guide developed for each question, the scorers assigned a score of 1 to 4, with 1 indicating no mastery and a 4 indicating complete mastery. Special codes were used to indicate responses that were: off topic (10), blank (11), or simply illegible (13). (Code 12, indicating a foreign language response, was not given to any of the responses.)

Table 1 shows the distribution of scores across students and across the six scores generated for each student. Two points are noteworthy. First, more than 10 percent of the papers were blank. We could not tell from available information whether a blank paper meant that the student did not have enough time to answer, was not motivated to answer, or simply did not know where to begin. We deleted blank, off topic, and illegible papers from our analyses, but in operational use such papers would be assigned a score of either zero or one. In any event, response rates for the essay questions were significantly different from those found earlier for the MC questions where nearly all students answered all or nearly all of the questions.

[1] Each ELA test booklet included two questions that required students to write an extended answer, usually several paragraphs. Question of this type are sometimes called open-ended, open-response, constructed-response, or extended constructed response questions. We refer to them as essay questions even though the responses are not always essays in the strictest sense.

The second noteworthy point about Table 1 is that relatively few students received full-credit for their responses. In fact, the number of responses scored as 4 was less than the number of blank responses. If blank papers are assigned a score of 0, then the average essay question score was just under 2.0. If blank papers are considered an indicator of low motivation and excluded, the average score for the remaining papers was 2.27.

Table 2.1
Distribution of Scores for the Essay Questions

Score	Number	Percent
1	4,668	20.7
2	6,780	30.0
3	6,298	27.9
4	1,806	8.0
Off Topic	668	3.0
Blank	2,377	10.5
Illegible	5	0.0

Difficulty. While minimum passing scores have not yet been established, it seems likely that students would have to score at least 3 to be considered as passing the standard measured by an essay question. If the purpose of the ELA portion of HSEE is to identify a relatively small proportion of students whose language arts or mathematics skills are below some minimally acceptable level of competency, then we would want relatively easy test questions. For essay questions, this would mean prompts and scoring guides where most students score in the 3 and 4 range, and only the very low-performing students would receive scores of 1 or 2. What we have is just the opposite with most students scoring 1 or 2 and only 8 percent receiving full credit. If, on the other hand, the intent is to require all students to perform above the current average (2.27 on the average essay question), then requiring a score of at least 3 for these essay questions just about right.

Regardless of the desired level of difficulty, it is essential that alternative questions have roughly the same level of difficulty. Each form will have a large number of multiple-choice questions and so parallel forms can be constructed by choosing a similar mix of easy and hard questions for each form. As currently envisioned, each form will have only two essay prompts, only one of which will be scored against the reading standards. There will not be much opportunity to balance easy and difficult questions, so the essay questions to be used in different forms must all have about the same difficulty.

The development contractor was highly successful in creating writing prompts and scoring guides of similar difficulty. Across the 24 prompts, the average scores ranged from 2.1 to 2.5, with most falling between 2.2 and 2.4. Scores were somewhat more variable across the 12 prompts scored for reading, ranging from 1.3 to 2.9. We would recommend dropping or revising 3 of the 12 reading prompts where half or more of all responses were assigned a score of 1. The minimum average for the remaining prompts was 1.9. If equal difficulty were the only goal, two prompts with average scores of 2.9 could also be dropped, leaving a range of score averages from 1.9 to 2.4. Overall, however, the prompts and scoring guides appear to be too difficult and so we would not recommend dropping the two easier prompts.

Item-Total Correlation. As with the MC questions, a second criterion in evaluating the quality of the essay questions is the extent to which scores on the question are consistent with information provided by all of the other questions. For MC questions, we looked for questions with an item-total score correlation[2] less than .20. Essay questions require a considerably greater investment of student time in responding and result in a maximum of 4 possible points compared to just 1 for multiple-choice questions. We expect more information from each essay score. Consequently, we chose to flag questions with item-total score correlations less than .40 as being inefficient.

Again, the writing scores were all highly efficient. Item-total correlations ranged between .53 and .73, all highly acceptable. The item-total correlation for one of the reading questions was .36, clearly below our cutoff. Item-total correlations for the remaining reading questions ranged from .42 to .60.

Disparate Impact. Table 2.2 shows average English/language arts MC and essay question scores for the different demographic groups typically included in disparate impact analyses. There were 100 MC questions, each scored 1 for correct responses or 0 for incorrect or omitted responses, so MC scores ranged from 0 to 100. Each student received three essay question scores, with each score ranging from 1 to 4, so the essay scores ranged from 3 to 12. Students who did not respond to both questions were excluded from these analyses. Table 2.2 also shows the standard deviation (SD) of the scores for each group. The standard deviation is a measure of how much the scores vary from the average. Roughly two-thirds of the scores will fall into the range running from one standard deviation below the average to one standard deviation above the average. The bottom half of Table 2.2 shows standardized differences. These are the difference between the average for a particular group and the overall average divided by the overall standard deviation. The purpose of this transformation is to provide comparisons across question formats that are adjusted for differences in the ranges of scores for these two formats.

The pattern of group differences for the essay questions is very similar to the pattern for the MC questions. More importantly, the differences among groups are not larger for the essay questions when converted to a common (standard deviation) metric.

[2] A correlation coefficient indicates the level of agreement between two measures. It ranges from -1.0 for perfect disagreement, where above-average scores on one measure are always accompanied by equally below-average scores on the second measure, to +1.0 for perfect agreement. The correlation coefficient will be 0.0 if there is no relationship between the two measures.

Table 2.2
Average Multiple Choice and Essay Scores by Demographic Group

Group	Multiple Choice (MC) (Percent Pass)			Essay Questions (Average Student Score)		
	Number	Average	SD	Number	Average	SD
ALL	3767	59.20	20.73	2997	6.94	2.13
Hispanic	1316	51.15	18.51	992	6.19	1.92
African American	197	50.51	20.22	138	6.33	2.06
Female	1840	63.26	19.14	1555	7.29	2.04
English language learners	430	40.73	14.60	265	5.62	1.78
Standardized Differences:						
Hispanic		-0.39	-0.11		-0.35	-0.10
Black		-0.42	-0.02		-0.29	-0.03
Female		0.20	-0.08		0.16	-0.04
English language learner		-0.89	-0.30		-0.62	-0.16

Table 2.3 provides very preliminary information on passing rates. To be sure, students will not pass or fail the essay questions separately; but scores on these questions will contribute to overall pass and fail decisions. It seems unlikely that performance on the essay questions will be considered satisfactory for students scoring 1 or 2 on the 4-point scale. We examined the effects of requiring an average score of 2.5 or a total score of 7.5 across the three essay scores for satisfactory performance. Overall, only 43 percent of the students would meet this criterion for the essay questions included in the Spring 2000 Field Test. The pass-rate for Hispanic and African American students would be less than 30 percent and the pass rates for students identified as English language learners would be less than 20 percent. Note also that the percentage of students not responding to one or both of the essay questions was significantly higher for the lower scoring groups. Overall, 80 percent of the students responded to both essay prompts. For African-American students, only 70 percent responded to both prompts and for English Language learners the figure was only about 60 percent.

Table 2.3
Percent with "Passing" Essay Question Scores by Demographic Group

Group	% Of Scores > 7.5 (If both essay questions answered)	% Missing One Essay	% Missing Both Essays
ALL	42.6	11.5	8.9
Hispanic	27.7	13.7	10.9
African American	29.7	15.7	14.2
Female	48.6	8.5	7.0
English language learners (ELL)	19.2	19.5	18.8

Differential Item Functioning. We used two relatively direct measures of differences across groups in the scores for each essay question. Other, more sophisticated indicators of group differences for multi-level scores have been identified (e.g., Zwick, Thayer & Mazzeo, 1997), but generally require larger sample sizes. Note that only about 300 students responded to each individual question. Across the 12 forms and subforms the number of students with valid responses to the essay questions (not blank, off-topic, or illegible) ranged from 242 to 341. The number of females ranged from 109 to 190 and the number of Hispanic students in each of these samples ranged from 76 to 121. The numbers for other demographic groups were generally less then 30, far too small for useful analyses.

First, we looked at group differences in average scores for each question relative to the average of these group differences across all questions. Across all of the questions, the average essay question score for Hispanic students was .2 less than the average score for all students. We flagged one reading question with a significantly greater difference (.46). All of the other questions had average score differences of .4 or less. No large differences were found by gender.

We examined group differences in item-total correlations as a second indicator of differential item functioning. The same reading question that showed a large mean difference for Hispanic students also had a significantly lower item-total correlation for these students (.36 compared to .59 for all students). All of the other differences in item-total correlations were less than .2. Also, all of the item-total correlations for Hispanic students were well above zero, indicating that the essay questions did function effectively for these students.

Statistical Screening Summary. All of the writing-only prompts passed all of the item screens. Writing scores for the dual use prompts also passed all of the screens. Table 2.4 summarizes the number of reading essay questions flagged for different statistical reasons. Overall 5 of the 12 reading questions were flagged. Editorial review may suggest that some of these questions are perfectly valid, so this represents a worst-case scenario. Overall the survival rate (percent of questions not flagged) for the writing questions was exceptional and the survival rate for reading was above 50%. It is quite common to find significantly lower survival rates in other similar programs. Note, however, that the statistical criteria for screening these questions were limited by sample size. We could not, for example, examine differential item functioning for African-American students, students with disabilities, or English language learners.

Table 2.4

Summary of Item Screening Results: Essay Reading Questions

Statistical Screen	Number Flagged	Booklet Number(s)
Low Passing Rates	3	3.2, 4.2, 4.3
Low Item-Total Correlation	1	1.3
DIF: Passing Rates	1	1.2
DIF: Item-Total Correlations	1	1.2
Total Flagged	5	1.2, 1.3, 3.2, 4.2, 4.3

Rater Agreement in Scoring the Essay Questions

Each essay was scored by two independent raters. Table 2.5 indicates the level of agreement of the two raters for each response. Entries in Table 2.5 show the number of papers receiving each possible combination of scores from the two independent raters across all of the students and essay questions. Counts on the diagonal of this table indicated the number of times the two raters gave the same score. In most cases where the two raters gave different scores, the scores were in adjacent categories, meaning that they differed by only one score point. The level of agreement is quite high as summarized in Table 2.6, which shows agreement level, by type of prompt and overall. There were, however, a small number of very dramatic differences where one rater assigned a score of 1 while the other assigned a score of 4. In an operational program, there is usually an "adjudication" process where disagreements of more than one score point are resolved by a third, typically more senior, rater.

Table 2.5
Counts of Essay Scores Assigned by Each Rater

Score assigned by the 1st rater	Score assigned by the 2nd rater							
	Valid Responses				Invalid Responses			Total
	1	2	3	4	Off-Topic	Blank	Illegible	
1	2045	285	11	6	1	0	0	2348
2	258	2704	371	16	0	2	0	3351
3	12	416	2605	104	0	0	0	3137
4	3	22	173	741	0	0	0	939
Off-Topic	2	2	0	0	318	8	0	330
Blank	0	0	1	0	16	1175	0	1192
Illegible	0	0	0	0	3	0	1	4
Total	2320	3429	3161	867	338	1185	1	11301

Table 2.6
Percent Agreement on Valid Responses by Question Type

Score Discrepancy	Type of Essay Question			
	Dual Score Prompts		Writing Only	All Prompts
	Reading	Writing		
Exact Match	84.9 %	81.0 %	82.5 %	82.8 %
1-Category	13.3 %	18.8 %	17.2 %	16.4 %
2+ Categories	1.8 %	0.1%	0.2 %	0.7 %
Total	100.0 %	100.0 %	100.0 %	100.0 %

We conducted a generalizability analysis (see Shavelson & Webb, 1991) as a final indication of the impact of discrepancies across scorers. In generalizability analyses, an estimate of true variation in student scores is compared to variation across different questions and scorers. The resulting information provides a basis for estimating the reliability of scores for different numbers of questions and scorers. We conducted a 3-question by 2-rater analysis of variance (see Scheffe, 1959) for each of the 12 form/subform combinations[3]. In these analyses, score variation by student is what we are trying to measure and so is labeled as "true" variance. The remaining sources of variation in scores are considered "error." Differences in the average score for each question is a source of error that will be eliminated through test form equating analyses. Interaction terms, such as student by question (S*Q), indicate the extent to which some students score higher on some questions while other

[3] We could not tell from available data exactly how many different scorers there were for each question or the extent to which the same or different scorers were used for different questions. We ran a variety of analyses with different assumptions about how scorers were nested within questions or students. Estimates for different sources of error varied slightly across these analyses, but the overall reliability estimates were essentially the same.

students score higher on other questions. Interactions between students and raters and between raters and questions are defined similarly. All are sources of error in the measurement of overall student achievement levels. Table 2.7 shows estimates for these different sources of score variation averaged across these 12 analyses.

Table 2.7

Sources of Variation in Scores for Essay Questions

Source	Type	Score Variance	% of Total
Student (S)	True	0.354	46.9 %
Question (Q)	Ignored[1]	0.102	N/A
Rater (R)	Error	0.000	0.0 %
S*Q	Error	0.303	40.2 %
S*R	Error	0.013	1.7%
Q*R	Error	0.000	0.0 %
S*Q*R	Error	0.083	11.0 %
TOTAL	True+Error	0.754	100 %

[1] Differences in question difficulties (main effects due to question) will be eliminated through equating.

Table 2.8 shows the “design” portion of the analyses, estimating the reliability for different numbers of questions and scorers. Reliability is a measure of score accuracy. It is equal to the ratio of "true" variation to the total variation in scores. When each rating of each response was considered separately, 47 percent of the total variation was "true" (between-student) variance so the reliability of a score from a single question would be .47. In the design analyses, statistical formulae are applied to estimate the reliability of scores that are averages across more than one question and/or more than one rater.

The overall reliability estimates are high considering that only three essay scores are included. Overall reliability, combining both MC and essay scores, is consistently in the range of .96. The proposed design of using two prompts to generate 3 scores increases the reliability of the essay question scores considerably in comparison to a single score from a single prompt. Adding a second rater does not increase the overall reliability very much. We do not, however, recommend using only a single scorer for each response. Because of the high-stakes nature of the individual student scores, a process for identifying and eliminating inconsistencies in scoring essay responses will be important. The few cases where one rater assigned a score of 1 while the other assigned a score of 4 illustrate, dramatically, the need for identifying (through multiple raters) and resolving (through a third reading) score discrepancies.

Table 2.8

Estimated Score Reliability by Number of Questions and Raters

Number of Questions	Number of Raters	
	1	2
1	.47	.50
2	.63	.66
3	72	.74

Revised Estimates of Test Accuracy

Overall, the accuracy of scores from the HSEE questions is likely to be quite high. The reliability estimate of .96 for ELA total scores means that the amount of measurement error is small (4 percent of the total score variation). This figure is quite good in comparison to most standardized tests. Reliabilities greater than .80 are considered acceptable for many purposes. For high-stakes uses, reliability estimates of .90 or higher are more commonly required.

Even with very high overall reliabilities, there will still be some inaccuracy in making pass-fail decisions based on a single score. Our June 30 report included extensive detail on analyses of the potential accuracy of HSEE total scores for ELA and mathematics when used to classify students as passing or failing. In this section we report the results of further analyses for ELA scores when essay scores are included in the total. We used an item response theory (IRT) model for multi-level scores, the Partial Credit Model (Muraki, 1992), to predict the distribution of scores for students resembling the field test participants. (See Wise, et al., June 30, for details on the procedures used.)

The new ELA scores include 100 MC questions plus 3 essay scores with 4 points each for a total of 112 possible scores. In these analyses, we assumed that blank and off-topic responses would be assigned a score of 0. Table 2.9 shows estimates of the percent of students who would score at different levels defined by plausible passing cutoffs. Again, we identified 50%, 60%, and 70% of the total possible score as plausible points for setting the minimum passing score. The addition of the essay scores leads to lower plausible passing rates in comparison to the prior analysis based on MC only. This difference should be interpreted cautiously, however, as higher omit rates for the essay questions may indicate lower effort on responses to these questions in this field-test setting.

Table 2.9
Number of Simulated Examinees at Different ELA Total Score Levels

Score Range	Minimum % Correct	Estimate % of Students	Estimated % Passing
0-55	0	38.0 %	
56-67	50	17.5%	62.0%
68-78	60	15.0 %	44.5%
79 – 112	70	29.5%	29.5%

There will, of course, always be some students whose true achievement level is right at the border between passing and failing. No test, no matter how reliable, can provide perfect classification for these students. To get an operational idea of what "near the border" might mean, we estimated the conditional standard errors (the standard error of measurement for students with a particular true number right score). Near the middle of the score range, these standard errors were 4.9 score points. To illustrate classification accuracy for students of different true achievement levels, we used the conditional standard error estimates to define a zone of uncertainty where student’s true achievement was very near the pass-fail border. Table 2.10 shows the number of students expected to have true achievement scores more than 5 points below or above a minimum score of 56.0. For each true achievement level, we estimated the number of students whose observed score from a single testing would be above or below 56. We classified these results as either correct or incorrect classifications, depending on whether the observed score level agreed with the examinee's true score. The results indicate that most students (more than 85%) would be clearly above or below the minimum and, for these students classification accuracy would be very high (98%–99%). For the 14.5% of students who are very near the minimum score level, about 70% (64% for those just below the passing level and 74 % for those just above) would be classified correctly.

Table 2.10
Estimated Percent Scoring Below/Above 56 Score Points by True Score Level

Subject	True (Expected) Number Correct	Percent of all Students	Percent of These Students Who Would Actually Score:	
			< 56	56+
ELA	00.00–51.99	30.5	97.8	2.2
	52.00-55.99	7.5	64.1	35.9
	56.00–59.99	7.0	26.6	73.4
	60.00+	55.0	0.8	99.2

Characteristics of the Field Test Samples

One important question is how well the students who participated in the field test represented the population of 10th grade public school students in California. AIR used 1999 STAR data to select representative samples of 100 schools each for the mathematics and English/language arts field tests. They then hoped to test 66 students from each school. The actual student participation rate varied considerably across schools and it is possible that more students participated from high (or low) performing schools than from low (or high) performing schools leading to a bias in estimates of student achievement levels. We conducted additional analyses to determine the extent to which this might be the case.

Table 2.11 shows a comparison of 10th grade STAR scores from spring 2000 for all schools in California and for schools participating in the HSEE field tests. Averages for all schools were weighted by the number of 10th graders in each school to generate averages for all students. This was the target against which results for the school and student samples were compared. Estimates for the schools in each of the field test samples were generated in two ways. First, the simple average of the school means was computed. This reflects the representativeness that would have resulted if the same number of students were tested from each school. Second, the means for each school were weighted by the number of field test participants from that school to provide an estimate of the effects of differential participation across schools. The results indicate a close correspondence with statewide averages (the first row in the table). There was a slight tendency to over-represent above-average schools in the ELA sample and a slight tendency to under-represent schools at very high and very low levels in both samples. Overall, however, these effects are slight.

Table 2.11
Comparison of FT Examinees to Statewide Averages: STAR 2000 Means and Standard Deviations

Population/Sample	Mathematics Average	SD	Reading Average	SD
Statewide – Weighted[1]	698	16.0	691	16.7
ELA Sample Schools	698	15.6	692	17.4
ELA Schools – Weighted[2]	699	14.6	694	16.2
Math Sample Schools	697	14.5	692	16.3
Math Schools – Weighted[2]	699	14.0	694	15.2

[1] Average STAR scores for each school in the state were weighted by the number of students in the school to compute the average score for all students. The standard deviation (SD) column shows the standard deviation of school averages when these weights are used.
[2] Average STAR scores for each participating school were weighted by the number of participants from that school to estimate average STAR scores for all of the students in the field test sample.

Tables 2.12 and 2.13 show similar comparisons using the 1999 STAR data, including demographic information that was not yet available for the 2000 STAR data. The field test samples were quite similar to the statewide averages for STAR reading and math scores. The demographic comparisons, however, show some under-representation of schools with higher proportions of Hispanic students, particularly for the math sample. For the demographic variables, we also have the responses from each of the students participating in the field test. Estimates of the percentage of Hispanic students based on these responses agree closely with percentages estimated from the overall school percent. This suggests that the students tested in each school were representative of the school as a whole, at least in this one respect. The percentage of English language learner (ELL) students tested in each school were slightly lower than percentages estimated from overall school percents, suggesting that ELL students were slightly underrepresented in the students tested from each school.

Table 2.12
Comparison of FT Examinees to Statewide Averages: STAR 1999 Means and Standard Deviations

Population/Sample	Mathematics Average	SD	Reading Average	SD
Statewide – Weighted[1]	697	16.3	690	16.6
ELA Sample Schools	696	15.2	691	16.4
ELA Schools – Weighted[2]	688	13.9	692	15.8
Math Sample Schools	696	14.4	691	16.5
Math Schools – Weighted[2]	698	13.5	693	15.4

See footnotes for table 2.11.

Table 2.13
Comparison of FT Examinees to Statewide Averages: Key 1999 10th Grade Demographics

Population/Sample	% Hispanic	% English language learners
Statewide – Weighted[1]	39	16
ELA Sample Schools	34	15
ELA Schools – Weighted[2]	35	16
ELA - Students Tested	36	14
Math Sample Schools	30	12
Math Schools – Weighted[2]	29	12
Math - Students Tested	29	10

See footnotes for table 2.11.

Summary

Scores for the ELA essay questions were analyzed to determine the quality of these questions and their scores. Five of the 12 reading questions were flagged for one or more potential statistical problems. None of the writing questions were flagged. Note however, that responses are available for only about 300 students for each question. A consequence was that analyses of differential item functioning for different demographic groups were quite limited.

Scoring consistency was analyzed and found to be quite high. Psychometric results suggested that a single read of each response by scorers might provide sufficient accuracy since the essay scores constitute only a small portion of the total scores. A more elaborate process may still be called for, however, to minimize challenges to results for individual students who end up just below the passing level.

We estimated the accuracy of ELA test scores and found to it be quite similar to the estimate provided in our June 30 report. These estimates were based on simulations that involved a number of assumptions. After key decisions about scoring and reporting are made and an intact form is administered under operational conditions, estimates of score accuracy involving fewer assumptions can be computed.

The schools participating in the field test appeared to be closely representative of the state as a whole. Student participation rates did not seem to be related to school performance means in a way that would bias estimates from the field test sample. However, the impact of within-schools non-participation and also of student motivation could not be estimated from available data.